Online Writer Identification Using Fuzzy C-means Clustering of Character Prototypes

Guoxian Tan, Christian Viard-Gaudin, Alex Kot

Online Writer Identification Using Fuzzy C-means Clustering of Character Prototypes

Guo Xian Tan
Nanyang Technological University of Singapore
tanguoxian@pmail.ntu.edu.sg

Christian Viard-Gaudin
IRCCyN/UMR CNRS 6597 Ecole Polytechnique de l'Université de Nantes
christian.viard-gaudin@univ-nantes.fr

Alex C. Kot
Nanyang Technological University of Singapore
eackot@ntu.edu.sg

Abstract

New kinds of documents such as handwritten online documents are emerging, which are produced by digital devices such as Tablet PC, personal handheld devices or digital paper coupled with digital pens. The rapid increase in the number of such handwritten online documents leads to mounting pressure on finding innovative solutions towards faster processing, indexing and retrieval of the documents from databases. One such method is to extract writer information derived from the raw ink signal for indexing and retrieval of the documents. This paper proposes a text independent method that does not place any constraints on the content being written or writing styles of the writers. We subsequently extract writer information at the character level from online handwritten documents and present a fuzzy c-means approach to cluster and classify the character prototypes for writer identification. The proposed system attained an accuracy of 97.6% on 82 writers and an accuracy of 98.3% when retrieved from a scaled up larger database of 120 writers.

Keywords: Writer identification, information retrieval, online handwriting, fuzzy c-means

1. Introduction

Technology has become an integral part of modern lifestyles that our lives are becoming intertwined with technology itself. Numerous initiatives have been funded to research and develop more efficient algorithms, software and computing platforms to handle the surge in demand for seamless interaction and to deliver interactive environments with a new level of intelligence [1, 2]. All these led to the emergence and proliferation of a kind of document: handwritten on-line documents. They are produced by state-of-the-art devices such as Tablet PC, personal handheld devices or digital paper coupled with digital stylus pens [3]. The rapid increase in the number of such documents requires efficient management tools for proper indexing and retrieving from databases.

Online handwritten digital documents are defined as those digital documents that not only provide information obtainable from offline digital documents, but also contain temporal information of the handwriting process [4]. Such additional information provides vital clues as to the identities of the writer. Writer identification systems must be clearly distinguished from writer verification systems. Writer verification performs a one-to-one matching between a test writer and a database of writers and attempts to ascertain the authenticity of the test writer. On the other hand, writer identification involves executing a one-to-many match and returns a ranked list of results for the search. The difference, though subtle, lies in the applications in which they can be utilized in.

Online document indexing using writer information provides two-fold distinct advantages. Firstly, from information security's point of view, writer identification has ubiquitous applications in digital rights management and forensic analysis in the prevention of fraud and identity theft cases. Secondly, in environments where large amounts of documents, forms, notes and meeting minutes are constantly being processed and managed, knowing the identity of the writer would provide an additional value. One such application is to process and retrieve the identities of students for subsequent verification purposes.

Online writer recognition systems can make use of global features such as texture, curvature and slant features [5, 6] as well as a combination of local features such as graphemes, allographs[1] and connected components [7, 8]. They can be generally classified into text-dependent or text-independent techniques. Previous works for online writer recognition such as the method proposed by Pitak et al. [9] adopted a Fourier transformation approach. The extracted features are the velocities of the barycenter of the pen movements and they are transformed into the frequency domain using Fourier transform. The advantage in adopting such a model is that it is text-independent, but

[1] Allographs are different shapes and forms of the same alphabet.

at the expense of a lower noise tolerance. The noise must be filtered out as much as possible in the pre-processing stage, otherwise the noise might be mistaken for high velocity components once the features are transformed into the frequency domain. Bensefia et al. [10, 11] proposed using a sequential clustering approach at the grapheme level to categorize different writers for their writer identification system. This approach attained an identification rate of 86% on a test set of 150 writers. The advantage of this method is that it does not depend on any lexicon and is therefore language independent. Works using allograph based methods are also popular in writer identification. Niels et al. used a dynamic time warping approach [4, 19, 20] to hierarchically cluster allographs and build a set of membership vectors, which contains the frequency of occurrence of each allograph for each character. This prototypic template of membership vectors then represents the handwriting styles of the different writers. Niels et al. reported a top1 accuracy of 89% based on this method [20]. Chan et al. [12] also made use of a character prototype distribution to model the specific allographs used by a given writer. They managed to achieve a top1 accuracy of 95% based on this text-independent approach. Even though working at the character level as opposed to using the grapheme or word level appears to be quite challenging, character prototype approaches are able to produce a more consistent set of templates for writer identification. Our proposed work further improves upon the work done by Chan et al. by adopting a fuzzy c-means algorithm, resulting in significant improvements in the performance.

The remainder of this paper is organized as follows: Section 2 describes the proposed methodology and experimental setup. Preprocessing such as normalization and resampling and our proposed fuzzy c-means algorithm is discussed here. Section 3 then presents the experimental results. Finally, discussions and future areas to explore are given in section 4.

2. Proposed Writer Identification Method

The writer identification can be divided into three stages, namely the prototype training stage, the reference and test document labeling stage and finally, the classification stage. During the prototype training stage, prototypes built at the character level are trained using the IRONOFF database [13] of 16585 isolated French words written by 373 subjects. The purpose of this stage is to build a set of character prototypes using the 16585 isolated words to model the different allographs of the 26 Latin alphabets (‘a’ to ‘z’). Following this, in the document labeling stage, automatically segmented characters from a set of 82 French reference and 82 test documents (extended later to 120 reference and 120 test documents) are then mapped to the set of prototypes built previously and transformed into a distribution of frequency vectors. These reference and test documents that we collected belong to a separate dataset from the IRONOFF dataset. A separate dataset needed to be collected because the IRONOFF dataset contains only isolated words and hence are not representative of actual online documents. Finally, the frequency vectors are used for classification in order to identify the writer corresponding to the test document. A detailed account of each of the three stages is given in [12]. Figure 1 shows the block diagram of the entire writer identification process.

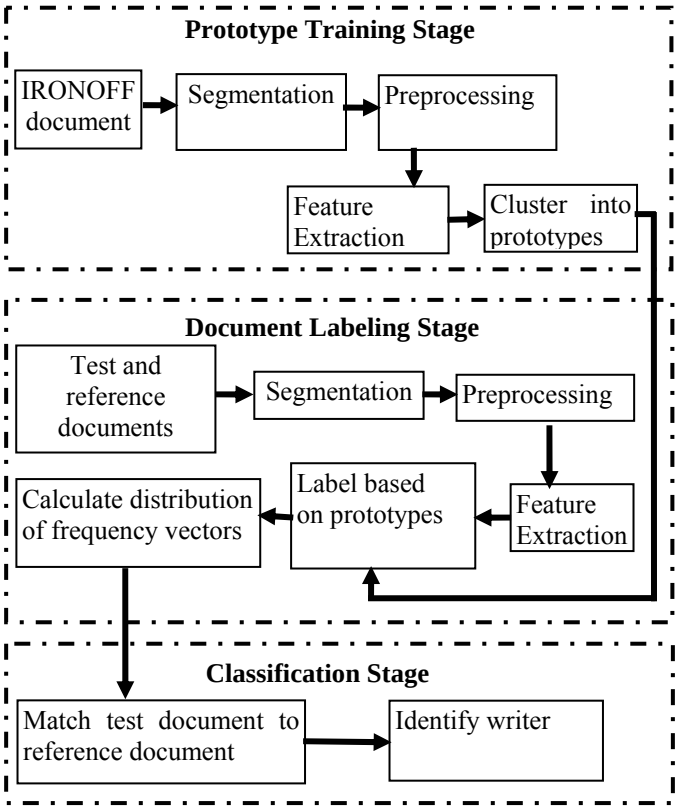

FIG. 1 – Block diagram for proposed methodology

One key point of the method relies on the automatic segmentation and labeling of the text, which is performed at the character level. This has to be done to extract and define the allographic prototypes from an independent word database (IRONOFF) and subsequently on all the reference documents and test documents to assess the writer retrieval performance of the proposed method. An industrial character segmentation and recognition engine, “MyScript SDK” [14], with the French linguistic resource attached for increased accuracy, has been used for this purpose. After the characters are segmented, the segmented characters then underwent further preprocessing where the size of each segmented character is normalized and resampled to 30 points. A process of feature extraction on each of the resampled points is then carried out. The features being used are the x and y co-

ordinates, the directions of x and y co-ordinates, the curvatures of x and y co-ordinates and the Pen-up or Pen-down information [12]. Accordingly, the matching process between a character extracted from a document and a prototype is carried in a feature space of dimension: 30 points × 7 features = 210.

In this paper, we propose a fuzzy c-means algorithm in the document labeling stage. The prototype training stage serves to identify common individual handwriting styles into individual prototypes at the character level. Following this, the document labeling stage then utilizes these prototypes to create individual distributions of handwriting styles for each of the test and reference documents in the database. Based on the results of the distributions, they provide statistical information about the handwriting styles of each writer. Our proposed method adopts a fuzzy c-means algorithm [15] which uses an exponential kernel function as described in Eq. 1 to create these individual distributions of handwriting styles.

$$C_{\alpha_{k_i}} = \sum_{p=1}^{M} \frac{\exp\left(-\beta \times dist(x_{\alpha_p}, x_{\alpha_k})\right)}{\sum_{k=1}^{N} \exp\left(-\beta \times dist(x_{\alpha_p}, x_{\alpha_k})\right)} \qquad (1)$$

where $C_{\alpha_{k_i}}$ is the total amount of characters from alphabet α, $\alpha \in \{'a','b',...,'z'\}$, that is assigned to prototype *k* from writer *i*. In Eq. 1, *p* represents a given segmented character that has been recognized as of the alphabet α, $p \in \{1,2,3,...,M\}$, with M being the number of characters corresponding to alphabet α. $dist(x_{\alpha_p}, x_{\alpha_k})$ is the Euclidean distance between the feature vector *x* for point *p* of alphabet α and the feature vector *x* for prototype k of alphabet α. In Eq. 1, β is a tuning parameter which is set to be 0.01 in our experiments.

Characters from the reference and test documents are then assigned a partial membership to the prototypes based on their distance metric to the prototypes. Therefore, characters which lie further away from certain prototypes are assigned a lesser degree to that particular prototype. $C_{\alpha_{k_i}}$ is then used to calculate the distribution of frequency vectors [12] to be used during classification.

It can be seen that a character does not need to be labeled to one particular prototype. There should be a fuzzy logic behind labeling the character to all the prototypes that are present and then building a distribution based on this fuzzy logic. The experimental results obtained serves as strong evidence to attest that our proposed fuzzy c-means approach for clustering into prototypes during the document labeling stage yielded a higher level of accuracy.

3. Experimental Results

The first set of experiments was initially conducted on a smaller set of 82 test and 82 reference documents from 82 different writers. As presented in table 1, the proposed methodology using fuzzy c-means to label the test and reference documents to the prototypes resulted in a high accuracy of 97.6% for writers that are ranked correctly in the top 1 position. This translates into a misclassification error of 2 misclassified writers out of 82, with both of them being misclassified in the top 2 position. This indicates that the writer identification system has confused the misclassified documents with only 1 other document from a different writer. Comparisons with previous results obtained by Siew et al. [12], who also performed writer identification based on character prototyping, show a significant improvement over their proposed methodology. An accuracy of 95.1% (4 misclassified writers out of 82) was reported in their method where the four writers were wrongly identified ranked at top 2, 4, 9, and 12 positions. This means that their writer identification system has confused the misclassified documents with up to 11 other documents. Therefore, our proposed methodology is able to perform with significantly higher accuracies.

Table 1. Performance of writer identification using different distance metrics

Size of reference document database	1NN[1]	Fuzzy C-means	
		Euclidean as distance metric	KL divergence[2] as distance metric
82	95.1%	97.6%	87.8%
120	96.7%	98.3%	91.7%

This improvement over Siew et al.'s results can be explained as follows. Their methodology hinges on the concept that each character can only be assigned to one particular prototype for which a distribution of handwriting styles is built. This is flawed in reality because there often exist overlapping handwriting styles for different writers. Our observations reveal that there are numerous instances when the characters are close to more than one prototype in the vector space. This can be explained by the fact that a writer can have strong, dominant handwriting style and weak handwriting styles. Weak handwriting styles change according to various circumstantial and temporal states [16], which can affect

[1] 1-Nearest Neighbor algorithm adopted by Siew et al. [12]
[2] Kullback-Leibler divergence

the strong dominant handwriting style and lead to reminiscence of multiple overlapping handwriting styles. Therefore, in our proposed methodology, each character is not just assigned to one particular prototype, but rather, each character is assigned a certain degree of all the prototypes depending on how close they are to that prototype. The more similar the character is to a certain prototype, the greater the degree that prototype has on the character. For example, one character might be attributed among three different prototypes if the character shares common traits of all three handwriting styles.

Experiments were also performed using the Kullback-Leibler (KL) divergence [17] as a different metric for the fuzzy c-means algorithm to determine the best performing metric for our writer identification system. We can observe from Table 1 that using KL divergence resulted in a low 87.8% identification rate. This can be attributed to the asymmetric nature of the KL divergence. Therefore, in our system, the better performing metric to use for our fuzzy c-means algorithm is the Euclidean distance metric.

The database was subsequently enlarged to 120 test and 120 reference documents from 120 different writers. The purpose of this experiment was to investigate the scalability of the system as to whether it can handle larger loads on an enlarged database. Table 1 show that a top 1 accuracy of 98.3% (2 misclassified writers out of 120) was obtained with both the remaining two wrongly classified writers at the top 2 positions. This indicates that the proposed methodology is indeed scalable and can handle applications where scalability is critical in designing the system.

3.1. Fuzzy C-means Kernel Design

Experiments were also conducted using different kernel functions [18] for the fuzzy c-means algorithm to determine the kernel function that can perform best in our writer identification system.

1. Gaussian kernel function: The distribution of the feature vectors was assumed to be Gaussian with zero mean and unit variance.

2. Inverse kernel function: Eq.2 describes the formulation for the inverse kernel function where the notations have the same meaning as that shown in Eq. 1.

$$C_{\alpha_{k_i}} = \sum_{p=1}^{M} \frac{\dfrac{1}{dist(x_{\alpha_p}, x_{\alpha_k})}}{\sum_{k=1}^{N} \dfrac{1}{dist(x_{\alpha_p}, x_{\alpha_k})}} \qquad (2)$$

Table 2 shows the comparison in performance of the writer identification among using the three different kernel functions. It can be seen that the inverse kernel function performs poorly, which can be explained by the poor behavior of such kernel functions as it approaches the centroids of the prototypes. The results obtained using the Gaussian kernel function is similar to that for the exponential kernel function. We believe that adjustments in the mean and variance parameters will allow the Gaussian kernel function to achieve the same results as the exponential kernel function since Gaussian functions are essentially exponential in nature.

Table 2. Performance of the Fuzzy c-means algorithm using different kernel functions.

Identification rate using exponential kernel function (Eq. 1)	98.3%
Identification rate using Gaussian kernel function	97.5%
Identification rate using inverse kernel function (Eq. 2)	96.7%

3.2. Effect of number of Character Prototypes on Accuracy

It is hypothesized that different alphabets require different number of prototypes to effectively model all the possible writing styles of that character. For example, there are more ways and styles to write the alphabet ‘f’ than to write the alphabet ‘c’. A preliminary level of analysis has been performed to find a global optimal value for the number of prototypes needed.

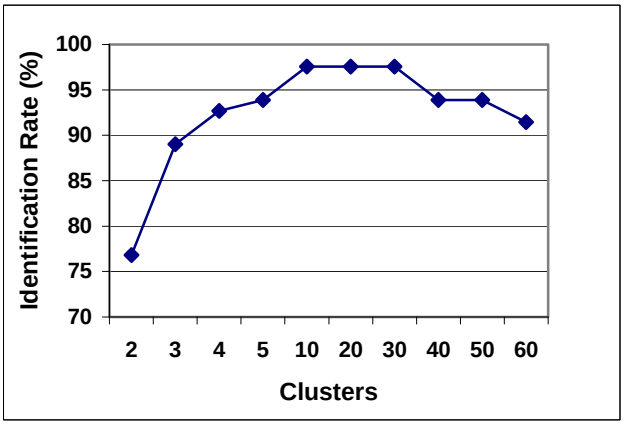

Figure 2. Graph of Identification Rate against Number of Clusters

Figure 2 shows the achievable by varying the number of global prototypes used for every letter of the alphabet. As seen from figure 2, the identification rate is highest when the number of prototypes varies from 10 to 30.

Additional number of prototypes beyond the 30 prototypes will result in a drop in the performance of the identification system. This can be explained by the principle of Occam's razor. A large number of prototypes create sparse dimensionality which deteriorates the performance of the classification. Likewise, insufficient number of prototypes will be unable to effectively separate between intra-class variations. Based on the above analysis, the optimum number of global prototypes is taken to be 10 in the experiment.

4. Discussions

From the experimental results, the proposed methodology is able to generate high accuracies of 97.6% (two misclassified writers out of 82) for the identification of 82 different writers, with both the misclassified writers being identified correctly in the rank 2nd position. This is a remarkable improvement over Siew et al.'s [12] methodology which only attained a top 1 accuracy of 95.12% (four misclassified writers out of 82), with all the misclassified writers only being correctly identified in the rank 11th position. This can be explained by the fact that our proposed methodology is based on fuzzy logic that mimics the fact that a writer's handwriting style has reminiscence of multiple overlapping handwriting styles. Furthermore, when we further increase the size of the reference database to 120 different writers, a similar result of having only two misclassified writers out of 120 was attained, leading to a top 1 accuracy of 98.3%. This concludes that the system can provide high accuracies in identifying writers and is highly scalable as well. Our results also indicated that KL divergence performed poorly in our writer identification system. This might be due to the fact that KL divergence itself is asymmetric. Furthermore, we have also showed that the optimum number of prototypes to use for our writer identification system is 10.

One main area of improvement to be explored is to investigate how different alphabets affect the character prototypes. Certain alphabets have more writing styles than others and are more discriminatory in writer identification. For instance, character shapes like 'f', 'h' or 't' have more variability than others like 'e' or 'c'. In this paper, we have presented the global optimum number of prototypes to use. However, we can go one step further in future to investigate a variable number of prototypes required to effectively model each alphabet, which may yield even better performance. Furthermore, more work needs to be done to investigate the extent of impact that different alphabets have on the accuracy of writer identification. In this way, a dynamic algorithm can then be proposed to handle and create an alphabet matrix for each writer. The alphabet matrix will be constructed based on the discriminative power of different alphabets for each writer and will then determine if certain alphabets can be ignored for writer identification. This will result in a smaller feature vector set being used, which will thus reduce the computational complexity of the system.

Acknowledgements

This research is jointly supported by Nanyang Technological University of Singapore, the French Merlion Scholarship and the ANR grant CIEL 06-TLOG-009.

References

[1] Moblile PC User Experience Guidelines for Developers, *http://msdn2.microsoft.com/en-us/library/ms695565(VS.85).aspx*, 2007.

[2] S. Oviatt, P. Cohen, L. Wu, L. Duncan, B. Suhm, J. Bers, T. Holzman, T. Winograd, J. Landay, J. Larson and D. Ferro, "Designing the User Interface for Multimodal Speech and Pen-Based Gesture Applications: State-of-the-Art Systems and Future Research Directions", *Human-Computer Interaction*, 2000, vol. 15, no. 4, pp. 263-322.

[3] Mobility and Tablet PC Road Map, *http://msdn2.microsoft.com/en-us/library/aa480233.aspx*, Jan 2007.

[4] A.K. Jain and A.M. Namboodiri, "Indexing and Retrieval of On-line Handwritten Documents", *Proceedings of the 7th International Conference on Document Analysis and Recognition*, 2003, pp. 655-659.

[5] A. Busch, W.W. Boles and S. Sridharan, "Texture for Script Identification", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 27, no. 11, Nov 2005, pp. 1720-1732.

[6] J. Hochberg, P. Kelly, T. Thomas and L. Kerns, "Automatic script identification from document images using cluster-based templates", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 19, no.2, Feb 1997, pp. 176-181.

[7] L. Schomaker and M. Bulacu, "Automatic Writer Identification Using Connected-Component Contours and Edge-Based Features of Uppercase Western Script", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, no. 6, Jun 2004, pp. 787-798.

[8] H. Srinivasan, S. Kabra, S. Srihari and C. Huang, "On Computing Strength of Evidence for Writer Verification", *Proceedings of the 9th International Conference on Document Analysis and Recognition*, Sep 2007, pp. 844-848.

[9] T. Pitak and T. Matsuura, "On-line Writer Recognition for Thai Based on Velocity of Bary center of Pen-point Movement", *Proceedings of IEEE International Conference on Image Processing*, Oct 2004, pp. 889-892.

[10] A. Bensefia, T. Paquet and L. Heutte, "A Writer Identification and Verification System", *Pattern Recognition Letters*, vol. 26, no. 13, 2005, pp. 2080-2092.

[11] A. Bensefia, T. Paquet and L. Heutte, "Information Retrieval Based Writer Identification", *Proceedings of the 7th International Conference on Document Analysis and Recognition*, 2003, pp. 946-950.

[12] S.K. Chan, C. Viard-Gaudin and Y.H. Tay "Online Text Independent Writer Identification Using Character Prototypes Distribution", *Proc. of SPIE-IS&T Electronic Imaging: Document Recognition and Retrieval XV*, 2008, vol. 6815, pp. 1-9.

[13] C. Viard-Gaudin, P-M Lallican, S. Knerr and P. Binter, "The IRESTE On/Off (IRONOFF) Dual Handwriting Database", *Proceedings of the 5th International Conference on Document Analysis & Recognition*, Sep 1999, pp. 455-458.

[14] Vision Objects Industrial Text Recogniser SDK, "MyScript Builder Help", SDK documentation, *http://www.visionobjects.com/about-us/download-center/_263/myscript-products-datasheets.html*, 2007.

[15] J. Han and M. Kamber, "*Data Mining: Concepts and Techniques*", Elsevier, 2006, pp. 383-460.

[16] R.A. Huber and A.M. Headrick, "*Handwriting Identification: Facts and Fundamentals*", CRC Press, 1999, pp. 87-139 and 175-243.

[17] T. Cover and J. Thomas, "*Elements of Information Theory"* Wiley, 1991, pp. 13-41.

[18] F. Hoppner, F. Klawonn, R. Kruse and T. Runkler, "*Fuzzy Cluster Analysis: Methods for Classification, Data Analysis and Image Recognition*", Wiley, 1999, pp. 5-31.

[19] R. Niels and L. Vuurpijl, "Generating Copybooks From Consistent Handwriting Styles", *Proceedings of the 9th International Conference on Document Analysis and Recognition*, Sep 2007, pp. 1009-1013.

[20] R. Niels, L. Vuurpijl and L. Schomaker, "Automatic Allograph Matching In Forensic Writer Identification", *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 21, no. 1, Feb 2007, pp. 61-81.